CENTRE@CLEF2018: Overview of the Replicability Task

Nicola Ferro[1], Maria Maistro[1], Tetsuya Sakai[2], and Ian Soboroff[3]

[1] Department of Information Engineering, University of Padua, Italy
{ferro,maistro}@dei.unipd.it
[2] Waseda University, Japan
tetsuyasakai@acm.org
[3] National Institute of Standards and Technology (NIST), USA
ian.soboroff@nist.gov

Abstract. Reproducibility has become increasingly important for many research areas, among those IR is not an exception and has started to be concerned with reproducibility and its impact on research results. This paper describes our first attempt to propose a lab on reproducibility named CENTRE and held during CLEF 2018. The aim of CENTRE is to run a reproducibility challenge across all the major IR evaluation campaigns and to provide the IR community with a venue where previous research results can be explored and discussed. This paper reports the participant results and preliminary considerations on the first edition of CENTRE@CLEF 2018, as well as some suggestions for future editions.

1 Introduction

Reproducibility is becoming a primary concern in many areas of science [16] as well as in computer science, as also witnessed by the recent ACM policy on result and artefact review and badging[4]. *Information Retrieval (IR)* is especially interested in reproducibility [12, 13, 34] since it is a discipline strongly rooted in experimentation where experimental evaluation represents a main driver of advancement and innovation.

Even if reproducibility has become part of the review forms at major conferences like SIGIR, this is more a qualitative assessment performed by a reviewer on the basis of what can be understood from a paper rather than an actual "proof" of the reproducibility of the experiments reported in the paper. Since 2015, the ECIR conference started a new track focused on reproducibility of previously published results. This conference track led to a stable enough flow of 3-4 reproducibility papers accepted each year but, unfortunately, this valuable effort did not produce a systematic approach to reproducibility: submitting authors adopted different notions of reproducibility, they adopted very diverse experimental protocols, they investigated the most disparate topics, resulting

[4] https://www.acm.org/publications/policies/artifact-review-badging

in a very fragmented picture of what was reproduced and what not, and the outcomes of these reproducibility papers are spread over a series of potentially disappearing repositories and Web sites.

Moreover, if we consider open source IR systems, they are typically used as:

- starting point by new-comers in the field, which take them almost off-the-shelf using default configuration to begin experience with IR and/or specific search tasks;
- base system on top of which to add a new component/technique you are interested to develop, keeping all the rest in the default configuration;
- baseline for comparison, again using default configuration.

Nevertheless, it has been repeatedly shown that best TREC systems still outperform off-the-shelf open source systems [2–4, 23, 24]. This is due to many different factors, among which lack of tuning on a specific collection when using default configuration, but it is also caused by the lack of the specific and advanced components and resources adopted by the best systems. It has been also shown that additivity is an issue, since adding a component on top of a weak or strong base does not produce the same level of gain [4, 23]. This poses a serious challenge when off-the-shelf open source systems are used as stepping stone to test a new component on top of them, because the gain might appear bigger starting from a weak baseline. Overall, the above considerations stress the need and urgency for a systematic approach to reproducibility in IR.

Therefore, the goal of CENTRE@CLEF 2018[5] is to run a joint task across CLEF/NTCIR/TREC on challenging participants:

- to reproduce best results of best/most interesting systems in previous editions of CLEF/NTCIR/TREC by using standard open source IR systems;
- to contribute back to the community the additional components and resources developed to reproduce the results in order to improve existing open source systems.

The paper is organized as follows: Section 2 introduces the setup of the lab; Section 3 discusses the participation and the experimental outcomes; and, Section 4 draws some conclusions and outlooks possible future works.

2 Evaluation Lab Setup

2.1 Tasks

The CENTRE@CLEF 2018 lab offered two pilot tasks:

- *Task 1 - Replicability*: the task focused on the replicability of selected methods on the same experimental collections;
- *Task 2 - Reproducibility*: the task focused on the reproducibility of selected methods on the different experimental collections.

[5] `http://www.centre-eval.org/clef2018/`

where we adopted the ACM Artifact Review and Badging definition of replicability and reproducibility:

- *Replicability (different team, same experimental setup)*: the measurement can be obtained with stated precision by a different team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same or a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using the author's own artifacts.
  In CENTRE@CLEF 2018 this meant to use the same collections, topics and ground-truth on which the methods and solutions have been developed and evaluated.
- *Reproducibility (different team, different experimental setup)*: The measurement can be obtained with stated precision by a different team, a different measuring system, in a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using artifacts which they develop completely independently.
  In CENTRE@CLEF 2018 this meant to use a different experimental collection, but in the same domain, from those used to originally develop and evaluate a solution.

2.2 Replicability and Reproducibility Targets

For this first edition of CENTRE we prefer to select the target runs among the Ad Hoc tasks in previous editions of CLEF, TREC, and NTCIR. We decided to focus on the Ad Hoc retrieval since it is a general and well known task. Moreover, the algorithms and the approaches used for Ad Hoc retrieval are widely used as basis for all the other types of tasks.

For each evaluation campaign, CLEF, TREC, and NTCIR, we considered all the Ad Hoc tracks and we examined the submitted papers and the proposed approaches. CLEF Ad Hoc tracks were proposed from 2000 to 2008, for TREC campaign we focused on the Web Track, from 2009 to 20011, while for NTCIR we checked the WWW track of 2017. To select the target papers among all the submitted ones we considered the following criteria:

- the popularity of the track, by accounting for the number of participating groups and the number of submitted runs;
- the impact of the proposed approach, measured by the number of citations received by the paper;
- the year of publication, since we preferred more recent papers;
- the tools used by the author, we discarded all those papers that were not using publicly available retrieval systems as Lucene, Terrier, Solr, and Indri.

Below we list the runs selected as targets of replicability and reproducibility among which the participants can choose. For each run, it is specified the

collection for replicability and the collections for reproducibility; for more information, the list also provides references to the papers describing those runs as well as the overviews describing the overall task and collections.

Since these runs were not originally thought for being used as targets of a replicability/reproducibility exercise, we contacted the authors of the papers to inform them and ask their consent to use the runs.

- **Run**: `AUTOEN` [18]
  - **Task type**: CLEF Ad Hoc Multilingual Task
  - **Replicability**: Multi-8 Two Years On with topics of CLEF 2005 [11]
  - **Reproducibility**: Multi-8 with topics of CLEF 2003 [29, 6]
- **Run**: `AH-TEL-BILI-X2EN-CLEF2008.TWENTE.FCW` [27]
  - **Task type**: CLEF Ad Hoc, Bilingual Task
  - **Replicability**: TEL English (BL) with topics of CLEF 2008 [1]
  - **Reproducibility**: TEL French (BNF) and TEL German (ONB) with topics of CLEF 2008 [1]
    TEL English (BL), TEL French (BNF) and TEL German (ONB) with topics of CLEF 2009 [15]
- **Run**: `AH-TEL-BILI-X2DE-CLEF2008.KARLSRUHE.AIFB_ONB_EN` [30]
  - **Task type**: CLEF Ad Hoc, Bilingual Task
  - **Replicability**: TEL German (ONB) with topics of CLEF 2008 [1]
  - **Reproducibility**: TEL English (BL) and TEL French (BNF) with topics of CLEF 2008 [1]
    TEL English (BL), TEL French (BNF) and TEL German (ONB) with topics of CLEF 2009 [15]

- **Run**: `UDInfolabWEB2` [33]
  - **Task type**: TREC Ad Hoc Web Task
  - **Replicability**: ClueWeb12 Category A with topics of TREC 2013 [9]
  - **Reproducibility**: ClueWeb09 Category A and B with topics of TREC 2012 [8]
    ClueWeb12 Category B with topics of TREC 2013 [9]
    ClueWeb12 Category A and B with topics of TREC 2014 [10]
- **Run**: `uogTrDwl` [26]
  - **Task type**: TREC Ad Hoc Web Task
  - **Replicability**: ClueWeb12 Category A with topics of TREC 2014 [10]
  - **Reproducibility**: ClueWeb09 Category A and B with topics of TREC 2012 [8]
    ClueWeb12 Category A and B with topics of TREC 2013 [9]
    ClueWeb12 Category B with topics of TREC 2014 [10]

- **Run**: `RMIT-E-NU-Own-1` and `RMIT-E-NU-Own-3` [17]
  - **Task type**: NTCIR Ad Hoc Web Task
  - **Replicability**: ClueWeb12 Category B with topics of NTCIR-13 [25]
  - **Reproducibility**: ClueWeb12 Category A with topics of NTCIR-13 [25]

The participants in CENTRE@CLEF 2018 were provided with the corpora necessary to perform the tasks. In details we made the following collections available:

- Multi-8 Two years On, is a document collection containing documents written in eight languages. A grand total of nearly 1.4 million documents in the languages Dutch, English, Finnish, French, German, Italian, Spanish and Swedish made up the multilingual collection;
- TEL data was provided by The European Library, the collection is divided in different subsets, each one corresponding to a different language: English, French and German, each of them containing about one million documents.
- ClueWeb09 consists of about 1 billion web pages in ten languages that were collected in January and February 2009. ClueWeb09 Category A represents the whole dataset, while ClueWeb09 Category B consists of the first English segment of Category A, which is roughly the first 50 million pages of the entire dataset;
- ClueWeb12 is the successor of ClueWeb2009 and consists of roughly 700 millions English web pages, collected between February 10, 2012 and May 10, 2012. ClueWeb12 Category A represents the whole dataset, while ClueWeb12 Category B is a uniform 7% sample of Category A.

Table 1 reports the number of documents and the languages of the documents contained in the provided corpora.

Table 1. Corpora used for the first edition of CENTRE@CLEF 2018 with the number of documents and the languages of the documents.

Name	Year	# Documents	Languages
Multi 8 Two Years On	2005	1,451,643	de, en, es, fi, fr, it, nl, sv
TEL English	2008	1,000,100	en
TEL French	2008	1,000,100	fr
TEL German	2008	869,353	de
ClueWeb09, Category A	2010	1,040,809,705	ar, de, en, es, fr, ja, ko, it, pt, zh
ClueWeb09, Category B	2010	50,220,423	en
ClueWeb12, Category A	2013	733,019,372	en
ClueWeb12, Category B	2013	52,343,021	en

Finally, Table 2 reports the topics used for the replicability and reproducibility tasks, with the corresponding number of documents, documents' languages and pool sizes. An example of topic for each evaluation campaign is reported in the Figure 1 for CLEF, 2 for TREC, and 3 for NTCIR.

2.3 Evaluation Measures

The quality of the replicability runs has been evaluated from two points of view:

Table 2. Topics used for the first edition of CENTRE@CLEF 2018 with the number of documents and the languages of the documents.

Evaluation Campaign	Task	# Topics	Languages
CLEF 2003	Multi-8	60	en, es
CLEF 2005	Multi-8 Two Years On	60	en, es
CLEF 2008	TEL bili X2DE	50	en, es, fr
CLEF 2008	TEL bili X2EN	153	de, es, fr,
CLEF 2008	TEL bili X2FR	50	de, en, es, nl
CLEF 2009	TEL bili X2DE	50	en, fr, it, zh
CLEF 2009	TEL bili X2EN	153	de, el, fr, it, zh
CLEF 2009	TEL bili X2FR	50	de, en, it
TREC 2012	Web Track, Ad Hoc Task	50	en
TREC 2013	Web Track, Ad Hoc Task	50	en
TREC 2014	Web Track, Ad Hoc Task	50	en
NTCIR-13	We Want Web Track	100	en

```
<topic lang="en">
    <identifier>456-AH</identifier>
    <title>Women's Vote in the USA</title>
    <description>Find publications on movements or actions aimed at obtaining voting rights
                 for women in the United States</description>
</topic>
<topic lang="fr">
    <identifier>456-AH</identifier>
    <title>Le vote des femmes aux États-Unis</title>
    <description>Trouver des documents sur les mouvements et les actions visant à obtenir
                 le droit de vote pour les femmes aux États-Unis.</description>
</topic>
```

Fig. 1. Example of a English topic with its French translation for CLEF 2008, task TEL bili X2DE.

```
<topic number="201" type="faceted">
    <query>raspberry pi</query>
    <description> What is a raspberry pi?</description>
    <subtopic number="1" type="inf">What is a raspberry pi?</subtopic>
    <subtopic number="2" type="inf">What software does a raspberry pi use?</subtopic>
    <subtopic number="3" type="inf">What are hardware options for a raspberry pi?</subtopic>
    <subtopic number="4" type="nav">How much does a basic raspberry pi cost?</subtopic>
    <subtopic number="5" type="inf">Find info about the raspberry pi foundation.</subtopic>
    <subtopic number="6" type="nav">Find a picture of a raspberry pi.</subtopic>
</topic>
```

Fig. 2. Example of topic for TREC 2013, Web Track, Ad Hoc Task.

```
<query>
    <qid>0008</qid>
    <content>World Table Tennis Championships</content>
</query>
```

Fig. 3. Example of topic for NTCIR-13, We Want Web Track.

– *Effectiveness*: how close are the performance scores of the replicated systems to those of the original ones. This is measured using the *Root Mean Square Error (RMSE)* [22] between the new and original *Average Precision (AP)* scores:

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(AP_{orig,i} - AP_{replica,i}\right)^2} \tag{1}$$

where m is the total number of topics, $AP_{orig,i}$ is the AP score of the original target run on topic t_i and $AP_{replica,i}$ is the AP score of the replicated run on topic t_i.

– *Ranked result lists*: since different result lists may produce the same effectiveness score, we also measure how close are the ranked results list of the replicated systems to those of the original ones. This is measured using Kendall's τ correlation coefficient [21] among the list of retrieved documents for each topic, averaged across all the topics. The Kendall's τ correlation coefficient on a single topic is given by:

$$\begin{aligned} \tau_i\big(orig, replica\big) &= \frac{P - Q}{\sqrt{\big(P + Q + T\big)\big(P + Q + U\big)}} \\ \bar{\tau}_i\big(orig, replica\big) &= \frac{1}{m}\sum_{i=1}^{m}\tau_i\big(orig, replica\big) \end{aligned} \tag{2}$$

where m is the total number of topics, P is the total number of concordant pairs (document pairs that are ranked in the same order in both vectors) Q the total number of discordant pairs (document pairs that are ranked in opposite order in the two vectors), T and U are the number of ties, respectively, in the first and in the second ranking.

Since for the reproducibility runs we do not have an already existing run to compare against, we planned to compare the reproduced run score with respect to a *baseline run* to see whether the improvement over the baseline is comparable between the original and the new dataset. However, we did not receive any reproducibility runs so we cannot put in practice this part of the evaluation task.

3 Participation and Outcomes

17 groups registered for participating in CENTRE@CLEF2018 but, unfortunately, only one group succeeded in submitting one replicability run.

Technical University of Wien (TUW) [19] replicated the run by Cimiano and Sorg, i.e. `AH-TEL-BILI-X2DE-CLEF2008.KARLSRUHE.AIFB_ONB_EN`. They submitted four runs described in Table 3, all the code they used to replicate the run is available online[6].

[6] `https://bitbucket.org/centre_eval/c2018_dataintelligence/src/master/`

Table 3. Submitted run files with their description.

Run Name	Description
esalength100_top10	ESA length: topic k = 1,000 records k = 100, top 10 documents
esalength1000_top10	ESA length: topic k = 10,000 records k = 1000, top 10 documents
esalength1000_top100	ESA length: topic k = 10,000 records k = 1000, top 100 documents
esalength1000_top1000	ESA length: topic k = 10,000 records k = 1000, top 1000 documents

The run AH-TEL-BILI-X2DE-CLEF2008.KARLSRUHE.AIFB_ONB_EN by Cimiano and Sorg uses a cutoff of 1,000 documents and so it has to be compared to esalength1000_top1000, which adopts the same cut-off. However, since TUW submitted runs also for cutoffs 10 and 100 documents, we compare them against versions of the run AH-TEL-BILI-X2DE-CLEF2008.KARLSRUHE.AIFB_ONB_EN capped at 10 and 100 documents per topic.

The paper by Cimiano and Sorg [30] uses *Cross-Lingual Explicit Semantic Analysis (CL-ESA)* to leverage Wikipedia articles to deal with multiple languages in a uniform way.

TUW encountered the following issues in replicating the original run:

- the Wikipedia underlying database dump of 2008 was no longer available and they have to resort to the static HTML dump of Wikipedia in the same period;
- the above issue caused a processing of Wikipedia articles sensibly different from the original one in [30] and had to rely on several heuristics to cope with HTML;
- they fixed an issue in the *Inverse Document Frequency (IDF)* computation, which might result in negative values according to the equation provided by [30];
- they had to deal with redirect pages in the static HTML dump of Wikipedia in order to find links across wiki pages in multiple languages;
- they had to find an alternative interpretation language identification heuristics.

All these issues prevented TUW from successfully replicating the original run. Indeed the *Mean Average Precision (MAP)* of the run by Cimiano and Sorg was 0.0667 while the MAP of the run esalength1000_top1000 by TUW is 0.0030.

The detailed results at different cutoffs for all the submitted runs are reported in table 4. It clearly emerges that all the above mentioned issues caused the TUW runs to severely underperform with respect to the original run and it is hard to say, in general, the extent to which it is possible to replicate it due to the changes in the language resources available.

The difficulties encountered in replicating the run are further stressed by the RMSE between AH-TEL-BILI-X2DE-CLEF2008.KARLSRUHE.AIFB_ONB_EN and esalength1000_top1000, computed according to eq. (1), which is 0.1132

Table 4. MAP at different cutoff thresholds $k = 10,\ 100,\ 1000$ for the original and the replicated runs.

	Original Run		Replicated Run
MAP@10	0.0121	0.0030	`esalength100_top10`
MAP@10	0.0121	0.0023	`esalength1000_top10`
MAP@100	0.0465	0.0029	`esalength1000_top100`
MAP@1000	0.0667	0.0030	`esalength1000_top1000`

and the average Kendall's τ correlation among the ranked lists of retrieved documents, computed according to eq. (2), which is $-5.69 \cdot 10^{-04}$.

Table 5 reports the results of the comparison with RMSE and Kendall's τ between the original and all the replicated runs. It can be noted how the RMSE deteriorates as the cutoff size increases as it is intuitive since it should be easier to stay closer to the original one when dealing with few top-k documents.

Table 5. RMSE and Kendall's τ between the original and the submitted runs

Run Name	RMSE	Kendall's τ
`esalength100_top10`	0.0282	-0.0363
`esalength1000_top10`	0.0283	-0.0474
`esalength1000_top100`	0.0879	-0.0090
`esalength1000_top1000`	0.1132	$-5.69 \cdot 10^{-04}$

Finally, Table 6 shows the Kendall's τ correlation between the submitted and the original runs computed for each single topic. We can observe as the general trend is to have very low correlations, i.e. very different document rankings between the topics, but there are a few exceptions. For example, topic `467-AH` has a correlation of 0.6644 with *Explicit Semantic Analysis (ESA)* length 1,000 and 10 documents cutoff, which suddenly drops to -0.0253 and -0.0038 for cutoffs 100 and 1,000, respectively, further stressing the fact that it should be easier to replicate the very top-k documents. Another interesting example is topic `490-AH` for which the correlation at ESA length 1,000 and 10 documents cutoff is -0.6170, indicating that the right documents have been retrieved but they have been ranked in almost reversed order.

4 Conclusions and Future Work

This paper reports the results on the first edition of CENTRE@CLEF2018. A total of 17 participants enrolled in the lab, however just one group managed to submit a run. As reported in the results section, the group encountered many

Table 6. Kendall's τ between the original and the submitted runs computed for each topic.

Topic	esalength100_top10	esalength1000_top10	esalength1000_top100	esalength1000_top1000
451-AH	0.4377	0.0829	0.0865	-0.0494
452-AH	-0.2722	-0.2916	-0.0621	-0.0371
453-AH	-0.4639	-0.1030	-0.1718	-0.0197
454-AH	0.2623	-0.3928	-0.0003	0.0310
455-AH	-0.0289	0.2416	0.0083	0.0150
456-AH	-0.0413	-0.1287	-0.0267	0.0047
457-AH	-0.6058	0.3882	0.0703	0.0106
458-AH	-0.0227	-0.4998	0.1372	0.0312
459-AH	-0.2459	-0.5021	0.0506	0.0337
460-AH	0.2945	-0.0990	-0.0068	0.0137
461-AH	-0.1637	-0.5951	-0.1277	0.0134
462-AH	-0.0266	-0.0180	0.0922	0.0219
463-AH	0.0827	0.1061	-0.0899	0.0303
464-AH	0.1069	-0.1421	-0.0651	0.0497
465-AH	-0.4081	0.1786	0.0257	-0.0182
466-AH	-0.3927	-0.2328	-0.0160	0.0153
467-AH	0.2807	0.6644	-0.0253	-0.0038
468-AH	-0.0755	-0.0951	-0.0496	0.0040
469-AH	-0.3870	-0.3166	0.1479	0.0283
470-AH	0.3339	-0.2227	-0.1398	-0.0076
471-AH	0.2954	0.0493	0.1211	0.0669
472-AH	-0.4701	-0.4440	-0.2120	-0.0160
473-AH	-0.2483	-0.1736	0.0142	-0.0441
474-AH	-0.2340	-0.5220	0.1072	0.0279
475-AH	-0.0475	0.2042	-0.0805	0.0316
476-AH	0.1730	-0.0754	0.1722	-0.0044
477-AH	-0.5510	-0.0394	0.0025	0.0479
478-AH	0.0264	0.5845	0.1340	0.0453
479-AH	0.4428	-0.0021	-0.0211	-0.0908
480-AH	-0.4202	0.0160	0.1304	-0.0682
481-AH	-0.4734	0.0248	-0.0674	0.0653
482-AH	0.0462	0.5746	-0.0264	0.0113
483-AH	-0.3460	0.0705	-0.0294	-0.0446
484-AH	0.4431	-0.3125	-0.0343	-0.0226
485-AH	-0.0411	-0.0088	-0.1511	-0.0502
486-AH	0.1177	-0.1333	-0.1147	-0.0528
487-AH	0.2000	0.2069	-0.2364	-0.0592
488-AH	-0.6633	0.3141	0.0783	-0.0360
489-AH	0.8133	0.0726	-0.0114	-0.0076
490-AH	-0.0426	-0.6170	0.1077	-0.0296
491-AH	-0.7382	-0.1575	-0.0343	-0.0045
492-AH	-0.4025	0.1594	-0.0021	0.0290
493-AH	0.1469	0.0091	-0.1079	0.0164
494-AH	0.2364	-0.4308	0.0432	0.0099
495-AH	0.2282	0.4465	0.0442	0.0340
496-AH	0.4478	0.2788	-0.0460	0.0291
497-AH	0.0108	-0.4886	-0.2636	-0.0602
498-AH	0.4813	-0.0073	0.1047	0.0251
499-AH	0.0701	0.1015	0.0759	0.0094
500-AH	0.0212	-0.0954	0.0158	-0.0539

substantial issues which prevented them to actually replicate the targeted run, as described in more detail in their paper [19].

These results support anecdotal evidence in the field about how much difficult is to actually replicate (and even more reproduce) research results, even in a field with such a long experimental tradition as IR is. However, the lack of participation is a signal that the community is somehow overlooking this important issue. As it also emerged from a recent survey within the SIGIR community [14], while there is a very positive attitude towards reproducibility and it is considered very important from a scientific point of view, there are many obstacles to it such as the effort required to put it into practice, the lack of rewards for achieving it, the possible barriers for new and inexperienced groups, and, least but not last, the (somehow optimistic) researcher's perception that their own research is already reproducible.

For the next edition of the lab we are planning to propose some changes in the lab organization to increase the interest and participation of the research community. First, we will target for newer and more popular systems to be reproduced, moreover we will consider other tasks than the AdHoc, as for example the medical or other popular domains.

References

1. Agirre, E., Di Nunzio, G.M., Ferro, N., Mandl, T., Peters, C.: CLEF 2008: Ad Hoc Track Overview. In: Peters, C., Deselaers, T., Ferro, N., Gonzalo, J., Jones, G.J.F., Kurimo, M., Mandl, T., Peñas, A. (eds.) Evaluating Systems for Multilingual and Multimodal Information Access: Ninth Workshop of the Cross–Language Evaluation Forum (CLEF 2008). Revised Selected Papers. pp. 15–37. Lecture Notes in Computer Science (LNCS) 5706, Springer, Heidelberg, Germany (2009)
2. Arguello, J., Crane, M., Diaz, F., Lin, J., Trotman, A.: Report on the SIGIR 2015 Workshop on Reproducibility, Inexplicability, and Generalizability of Results (RIGOR). SIGIR Forum 49(2), 107–116 (December 2015)
3. Armstrong, T.G., Moffat, A., Webber, W., Zobel, J.: Has Adhoc Retrieval Improved Since 1994? In: Allan, J., Aslam, J.A., Sanderson, M., Zhai, C., Zobel, J. (eds.) Proc. 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2009). pp. 692–693. ACM Press, New York, USA (2009)
4. Armstrong, T.G., Moffat, A., Webber, W., Zobel, J.: Improvements That Don't Add Up: Ad-Hoc Retrieval Results Since 1998. In: Cheung, D.W.L., Song, I.Y., Chu, W.W., Hu, X., Lin, J.J. (eds.) Proc. 18th International Conference on Information and Knowledge Management (CIKM 2009). pp. 601–610. ACM Press, New York, USA (2009)
5. Borri, F., Nardi, A., Peters, C., Ferro, N. (eds.): CLEF 2008 Working Notes. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, `http://ceur-ws.org/Vol-1174/` (2008)
6. Braschler, M.: CLEF 2003 – Overview of Results. In: Peters et al. [28], pp. 44–63
7. Braschler, M., Di Nunzio, G.M., Ferro, N., Peters, C.: CLEF 2004: Ad Hoc Track Overview and Results Analysis. In: Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) Multilingual Information Access for Text,

Speech and Images: Fifth Workshop of the Cross–Language Evaluation Forum (CLEF 2004) Revised Selected Papers. pp. 10–26. Lecture Notes in Computer Science (LNCS) 3491, Springer, Heidelberg, Germany (2005)
8. Clarke, C.L.A., Craswell, N., Voorhees, E.M.: Overview of the TREC 2012 Web Track. In: Voorhees, E.M., Buckland, L.P. (eds.) The Twenty-First Text REtrieval Conference Proceedings (TREC 2012). pp. 1–8. National Institute of Standards and Technology (NIST), Special Publication 500-298, Washington, USA (2013)
9. Collins-Thompson, K., Diaz, F., Clarke, C.L.A., Voorhees, E.M.: TREC 2013 Web Track Overview. In: Voorhees [31]
10. Collins-Thompson, K., Macdonald, C., Bennett, P.N., Voorhees, E.M.: TREC 2014 Web Track Overview. In: Voorhees and Ellis [32]
11. Di Nunzio, G.M., Ferro, N., Jones, G.J.F., Peters, C.: CLEF 2005: Ad Hoc Track Overview. In: Peters, C., Gey, F.C., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B., Müller, H., de Rijke, M. (eds.) Accessing Multilingual Information Repositories: Sixth Workshop of the Cross–Language Evaluation Forum (CLEF 2005). Revised Selected Papers. pp. 11–36. Lecture Notes in Computer Science (LNCS) 4022, Springer, Heidelberg, Germany (2006)
12. Ferro, N.: Reproducibility Challenges in Information Retrieval Evaluation. ACM Journal of Data and Information Quality (JDIQ) 8(2), 8:1–8:4 (February 2017)
13. Ferro, N., Fuhr, N., Järvelin, K., Kando, N., Lippold, M., Zobel, J.: Increasing Reproducibility in IR: Findings from the Dagstuhl Seminar on "Reproducibility of Data-Oriented Experiments in e-Science". SIGIR Forum 50(1), 68–82 (June 2016)
14. Ferro, N., Kelly, D.: SIGIR Initiative to Implement ACM Artifact Review and Badging. SIGIR Forum 52(1) (June 2018)
15. Ferro, N., Peters, C.: CLEF 2009 Ad Hoc Track Overview: TEL & Persian Tasks. In: Peters, C., Di Nunzio, G.M., Kurimo, M., Mandl, T., Mostefa, D., Peñas, A., Roda, G. (eds.) Multilingual Information Access Evaluation Vol. I Text Retrieval Experiments – Tenth Workshop of the Cross–Language Evaluation Forum (CLEF 2009). Revised Selected Papers. pp. 13–35. Lecture Notes in Computer Science (LNCS) 6241, Springer, Heidelberg, Germany (2010)
16. Freire, J., Fuhr, N., Rauber, A. (eds.): Report from Dagstuhl Seminar 16041: Reproducibility of Data-Oriented Experiments in e-Science. Dagstuhl Reports, Volume 6, Number 1, Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Germany (2016)
17. Gallagher, L., Mackenzie, J., Benham, R., Chen, R.C., Scholer, F., Culpepper, J.S.: RMIT at the NTCIR-13 We Want Web Task. In: Kando et al. [20], pp. 402–406
18. Guyot, J., Radhouani, S., Falquet, G.: Ontology-Based Multilingual Information Retrieval. In: Peters, C., Quochi, V., Ferro, N. (eds.) CLEF 2005 Working Notes. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, `http://ceur-ws.org/Vol-1171/` (2005)
19. Jungwirth, M., Hanbury, A.: Replicating an Experiment in Cross-lingual Information Retrieval with Explicit Semantic Analysis. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) CLEF 2018 Working Notes. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073 (2018)
20. Kando, N., Fujita, S., Kato, M.P., Manabe, T. (eds.): Proc. 13th NTCIR Conference on Evaluation of Information Access Technologies. National Institute of Informatics, Tokyo, Japan (2017)
21. Kendall, M.G.: Rank correlation methods. Griffin, Oxford, England (1948)
22. Kenney, J.F., Keeping, E.S.: Mathematics of Statistics – Part One. D. Van Nostrand Company, Princeton, USA, 3rd edn. (1954)

23. Kharazmi, S., Scholer, F., Vallet, D., Sanderson, M.: Examining Additivity and Weak Baselines. ACM Transactions on Information Systems (TOIS) 34(4), 23:1–23:18 (June 2016)
24. Lin, J., Crane, M., Trotman, A., Callan, J., Chattopadhyaya, I., Foley, J., Ingersoll, G., Macdonald, C., Vigna, S.: Toward Reproducible Baselines: The Open-Source IR Reproducibility Challenge. In: Ferro, N., Crestani, F., Moens, M.F., Mothe, J., Silvestri, F., Di Nunzio, G.M., Hauff, C., Silvello, G. (eds.) Advances in Information Retrieval. Proc. 38th European Conference on IR Research (ECIR 2016). pp. 357–368. Lecture Notes in Computer Science (LNCS) 9626, Springer, Heidelberg, Germany (2016)
25. Luo, C., Sakai, T., Liu, Y., Dou, Z., Xiong, C., Xu, J.: Overview of the NTCIR-13 We Want Web Task. In: Kando et al. [20], pp. 394–401
26. McCreadie, R., Deveaud, R., Albakour, D., Mackie, S., Limsopatham, N., Macdonald, C., Ounis, I., Thonet, T.: University of Glasgow at TREC 2014: Experiments with Terrier in Contextual Suggestion, Temporal Summarisation and Web Tracks. In: Voorhees and Ellis [32]
27. Nguyen, D., Overwijk, A., Hauff, C., Trieschnigg, D.R., Hiemstra, D., de Jong, F.M.G.: WikiTranslate: Query Translation for Cross-lingual Information Retrieval Using only Wikipedia. In: Borri et al. [5]
28. Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (eds.): Comparative Evaluation of Multilingual Information Access Systems: Fourth Workshop of the Cross–Language Evaluation Forum (CLEF 2003) Revised Selected Papers. Lecture Notes in Computer Science (LNCS) 3237, Springer, Heidelberg, Germany (2004)
29. Savoy, J.: Report on CLEF-2003 Multilingual Tracks. In: Peters et al. [28], pp. 64–73
30. Sorg, P., Cimiano, P.: Cross-lingual Information Retrieval with Explicit Semantic Analysis. In: Borri et al. [5]
31. Voorhees, E.M. (ed.): The Twenty-Second Text REtrieval Conference Proceedings (TREC 2013). National Institute of Standards and Technology (NIST), Special Publication 500-302, Washington, USA (2014)
32. Voorhees, E.M., Ellis, A. (eds.): The Twenty-Third Text REtrieval Conference Proceedings (TREC 2014). National Institute of Standards and Technology (NIST), Special Publication 500-308, Washington, USA (2015)
33. Yang, P., Fang, H.: Evaluating the Effectiveness of Axiomatic Approaches in Web Track. In: Voorhees [31]
34. Zobel, J., Webber, W., Sanderson, M., Moffat, A.: Principles for Robust Evaluation Infrastructure. In: Agosti, M., Ferro, N., Thanos, C. (eds.) Proc. Workshop on Data infrastructurEs for Supporting Information Retrieval Evaluation (DESIRE 2011). pp. 3–6. ACM Press, New York, USA (2011)